

Université d'Aix-Marseille

M1 MAS

Données catégorielles : Exploration, Modélisation

Analyse Factorielle des Correspondances (AFC)\*

Nicolas Pech  
nicolas.pech@univ-amu.fr  
04 13 55 07 73

Année 2022-2023

## Table des matières

<b>1</b>	<b>Tableau de contingence et profils</b>	<b>2</b>
<b>2</b>	<b>Ecart à l'indépendance</b>	<b>3</b>
2.1	Lois marginales . . . . .	4
2.2	Lois conditionnelles . . . . .	4
2.3	Mesure empirique d'écart à l'indépendance . . . . .	4
2.4	Test du khi-deux . . . . .	5
<b>3</b>	<b>Principes de l'AFC</b>	<b>8</b>
3.1	Analyse du nuage des profils-lignes . . . . .	8
3.2	Analyse du nuage des profils-colonnes . . . . .	9
3.3	Métrique du khi2 . . . . .	9
3.4	Inertie des nuages . . . . .	10
3.5	ACP des deux nuages profils . . . . .	11
3.6	Relations de transition . . . . .	12
3.7	Formules de reconstitution . . . . .	13
<b>4</b>	<b>Aides à l'interprétation</b>	<b>14</b>
<b>5</b>	<b>Interpréter les résultats d'une AFC</b>	<b>14</b>
<b>6</b>	<b>Références</b>	<b>21</b>

---

\*Ce document est largement inspiré d'une version précédente rédigée par M-C. Roubaud.

L'Analyse Factorielle des Correspondances (AFC) a été proposée en France dans les années soixante par J.P. Benzécri dans le but d'étudier la liaison (ou correspondance) entre deux variables qualitatives. Cette méthode permet l'analyse de tableaux de contingence obtenus en observant deux variables qualitatives sur un ensemble de  $n$  individus (unités statistiques) et certains tableaux binaires (tableaux dits de "présence-absence"). C'est une méthode très utilisée en analyse des données textuelles. L'Analyse Factorielle des Correspondances (AFC) peut être présentée comme une analyse en composantes principales avec une métrique spéciale, la métrique du khi-deux. C'est cette approche qui sera privilégiée ici, en utilisant les acquis du cours de statistique exploratoire.

### Exemples

- On observe sur 10 000 étudiants les deux variables qualitatives  $X$  et  $Y$  suivantes :  
 $X$  : CSP du père et  $Y$  : Type d'études supérieures poursuivies.  
→ A quoi correspond ici un individu ?
- On analyse les réponses de 2000 personnes à une question ouverte et on s'intéresse au tableau croissant les 21 mots (formes) les plus utilisés avec 5 niveaux de diplômes des personnes interrogées. Ici  $X$  correspond aux formes (variable qualitative avec 21 modalités) et  $Y$  aux diplômes (5 modalités).  
→ A quoi correspond ici un individu ?
- On considère l'observation pour chacun de 592 individus de deux variables : la couleur des yeux et la couleur des cheveux. Les observations sont présentées dans la table 1.

	brun	châtain	roux	blond	
marron	68	119	26	7	
noisette	15	54	14	10	
vert	5	29	14	16	
bleu	20	84	17	94	

TABLE 1 –

→ Décrire la table 1. Quel est l'intérêt de cette présentation des observations ? Y-en-a-t-il une autre ? En déduire la légende de la table 1

## 1 Tableau de contingence et profils

**Définition 1.1** Un tableau de contingence  $N$  est obtenu en croisant deux variables qualitatives :  $X$  à  $I$  modalités (notées  $x_1, \dots, x_I$ ) et  $Y$  à  $J$  modalités (notées  $y_1, \dots, y_J$ ) observées sur  $n$  individus. Le tableau  $N$  à  $I$  lignes et  $J$  colonnes a pour élément général :

$n_{ij} = \text{nbre d'individus sur lesquels on a observé les modalités } x_i \text{ de } X \text{ et } y_j \text{ de } Y.$

$X/Y$	$y_1$	$\cdots$	$y_j$	$\cdots$	$y_J$	total
$x_1$	$n_{11}$	$\cdots$	$n_{1j}$	$\cdots$	$n_{1J}$	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i$	$n_{i1}$	$\cdots$	$n_{ij}$	$\cdots$	$n_{iJ}$	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_I$	$n_{I1}$	$\cdots$	$n_{Ij}$	$\cdots$	$n_{IJ}$	$n_{I\cdot}$
total	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot J}$	$n$

L'opération consistant à établir un tel tableau est appelée un "tri croisé" dans le domaine de l'enquête.  
On posera

$$N = \begin{pmatrix} n_{11} & \dots & n_{1j} & \dots & n_{1J} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{I1} & \dots & n_{Ij} & \dots & n_{IJ} \end{pmatrix}$$

**Définition 1.2** On appelle *ième effectif marginal des lignes*  $n_{i\cdot} = \sum_{j=1}^J n_{ij}$ ,

On appelle *jème effectif marginal des colonnes*



Il existe deux lectures différentes du même tableau : en ligne ou en colonne.

**Définition 1.3** On appelle *tableau des profils-lignes* le tableau des fréquences conditionnelles  $\frac{n_{ij}}{n_{i\cdot}}$  (la somme de chaque ligne est ramenée à 1).

$$\text{ième profil-ligne} = \left( \frac{n_{i1}}{n_{i\cdot}}, \dots, \frac{n_{is}}{n_{i\cdot}} \right) \in \mathbb{R}^J$$

Le *ième profil-ligne* est la distribution empirique des modalités  $y_1, y_2, \dots, y_J$  parmi les individus possédant la modalité  $x_i$ .

→ Compléter la définition suivante :

**Définition 1.4** On appelle *tableau des profils-colonnes*

→ A quoi correspond chacune des deux matrices suivantes ?

→ Exprimer en une phrase la signification de chacune des valeurs indiquées en rouge

$$T_{pc} = \begin{pmatrix} 0.63 & 0.42 & 0.37 & .06 \\ 0.14 & 0.19 & \textcolor{red}{0.20} & 0.08 \\ 0.05 & 0.10 & 0.20 & 0.13 \\ 0.19 & 0.29 & 0.24 & 0.74 \end{pmatrix} \quad T_{pl} = \begin{pmatrix} .31 & .54 & .12 & .03 \\ .16 & .58 & \textcolor{red}{.15} & .11 \\ .08 & .45 & .22 & .25 \\ .09 & .39 & .08 & .44 \end{pmatrix}$$

## 2 Ecart à l'indépendance

**Définition 2.1** On appelle *loi de probabilité jointe* ou *loi du couple*  $(X, Y)$  la loi de probabilité  $P_{(X,Y)}$  définie ainsi :  $P_{(X,Y)}(x_i, y_j) = P(X = x_i \cap Y = y_j)$ .

Dans le cas discret, cette loi de probabilité peut se mettre sous la forme d'une table :

	$y_1$	$\dots$	$y_j$	$\dots$	$y_J$	
$x_1$				⋮		
$\vdots$				⋮		
$x_i$		$\dots$	$\dots$	$p_{ij}$	$\dots$	$\dots$
$\vdots$				⋮		
$x_I$						$p_{\cdot j}$

On note  $p_{ij} = P(X = x_i \cap Y = y_j)$ , et bien entendu :  $\sum_{i=1}^I \sum_{j=1}^J p_{ij} = \square$

## 2.1 Lois marginales

**Définition 2.2** On appelle *lois marginales*, les lois de probabilité de  $X$  et  $Y$  prises séparément :

$$P(X = x_i) = \sum_{j=1}^J p_{ij} = p_i.$$

$$P(Y = y_j) = \sum_{i=1}^I p_{ij} = p_{\cdot j}$$

## 2.2 Lois conditionnelles

Lorsque les événements  $\{X = x_i\}$  et  $\{Y = y_j\}$  sont de probabilités non nulles, on peut définir deux types de lois conditionnelles selon que l'on connaît la valeur de  $X$  ou de  $Y$ .

**Définition 2.3** On appelle *loi conditionnelle de  $X$  si  $Y = y_j$*  :

$$P(X = x_i / Y = y_j) = \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{\cdot j}}$$

On appelle *loi conditionnelle de  $Y$  si  $X = x_i$*  :

$$P(Y = y_j / X = x_i) = \frac{P(X = x_i \cap Y = y_j)}{P(X = x_i)} = \frac{p_{ij}}{p_i}$$

**Propriété 2.1** Si  $X$  et  $Y$  sont deux variables aléatoires indépendantes, alors :

$$p_{ij} = P(X = x_i \cap Y = y_j) = P(X = x_i) \times P(Y = y_j) = p_i \times p_{\cdot j} \quad \forall(i, j)$$

## 2.3 Mesure empirique d'écart à l'indépendance

L'indépendance empirique se traduit par :  $\frac{n_{ij}}{n} = \frac{n_i}{n} \frac{n_{\cdot j}}{n}$  c'est à dire  $n_{ij} = \frac{n_i \cdot n_{\cdot j}}{n}$ .

→ Dans le cas où l'indépendance empirique serait respectée, quelles caractéristiques auraient les données de l'exemple ?

Dans le cas général on peut définir une mesure de l'écart entre les observations et la situation d'indépendance empirique :

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \frac{n_i \cdot n_{\cdot j}}{n})^2}{\frac{n_i \cdot n_{\cdot j}}{n}}, \text{ où}$$

- $n_{ij}$  est l' effectif observé dans la case  $(i, j)$

- $\frac{n_{i,j} \cdot n_{\cdot,j}}{n}$  est

On note que dans le cas de l'indépendance :  $\chi^2 = \boxed{\quad}$ .

On montre que  $\chi^2 \leq n \times \min(I - 1, J - 1)$  et que la borne est atteinte dans le cas de la dépendance fonctionnelle entre  $X$  et  $Y$ .

→ La table 2 représente les effectifs correspondant à la situation d'indépendance de l'exemple des couleurs. Que représente alors la table 3 et que peut on en déduire ?

	brun	châtain	roux	blond	total
marron	40,14	106,284	26,39	47,20	220
noisette	16,97	44,93	11,15	19,95	93
vert	11,68	30,92	7,68	13,73	64
bleu	39,22	103,87	25,79	46,12	215
total	108	286	71	127	592

TABLE 2 – Tableau de contingence présentant la répartition de 592 personnes suivant leurs couleurs des yeux et des cheveux dans la situation d'indépendance entre les deux variables

	brun	châtain	roux	blond	total
marron	19,35	1,52	0,01	34,23	55,11
noisette	0,23	1,83	0,73	4,96	7,75
vert	3,82	0,12	5,21	0,382	9,52
bleu	9,42	3,80	2,99	49,70	65,91
total	32,81	7,27	8,94	89,27	138,29

TABLE 3 –

Pour chaque case  $(i, j)$  du tableau, il est important d'étudier sa contribution à  $\chi^2$  :

$$\frac{(\text{effectif observé} - \text{effectif attendu})^2}{\text{effectif attendu}}$$

On peut également calculer le pourcentage par rapport à  $\chi^2$ . Ceci met en évidence les associations significatives entre modalités des deux variables. Le signe de la différence (*effectif observé - effectif attendu*) indique s'il s'agit d'une association sur-représentée (positive) ou sous-représentée (négative).

D'autres indices ont été proposés pour mesurer la dépendance entre deux variables. On pourra voir par exemple le polycopié *La procédure FREQ de SAS : Tests d'indépendance et mesures d'association dans un tableau de contingence* de J. Confais, Y. Grelet et M. Le Guen ou encore celui de Ricco Rakotomalala : *Etude des dépendances - Variables qualitatives, Tableau de contingence et mesures d'association*, tous les deux sur la page AMeTICE du cours.

## 2.4 Test du khi-deux

A partir de l'observation de la statistique  $\chi^2$  on peut s'interroger sur la significativité de la liaison entre les deux variables. Répondre à cette question consiste à réaliser un test

- de  $(H_0)$  : *Les variables X et Y sont indépendantes*  
 contre  $(H_1)$  : *Les variables X et Y ne sont pas indépendantes*

- **Sous  $H_0$ ,**  $\chi^2$  est une réalisation d'une variable aléatoire  $Z^2$  suivant approximativement une loi du  $\chi^2_{(I-1)(J-1)}$ .
- **Règle usuelle d'application :**  $effectif\ attendu \geq 5$  dans toutes les cases. Nous considérerons que l'approximation est acceptable dès que 80% des cases ont un  $effectif\ attendu \geq 5$ .
- **Calcul de la p-valeur :**  
 La p-valeur est la probabilité d'obtenir une valeur au moins égale à la valeur observée  $\chi^2$  si les deux variables étaient indépendantes :

$$\text{p-valeur} = P_{H_0}(Z^2 > \chi^2)$$

Cette probabilité est lue dans la table de la loi du khi2.

- **Décision :** On choisit un risque de première espèce  $\alpha$ , par exemple  $\alpha = 1\%$  ou  $5\%$ .  
*Si p-valeur  $\leq \alpha$ , on rejette  $H_0$ , on décide donc de considérer les variables comme étant dépendantes.*  
*Si non, on ne peut rejeter  $H_0$  au risque  $\alpha$  et on décide donc de ne pas rejeter l'hypothèse d'indépendance.*

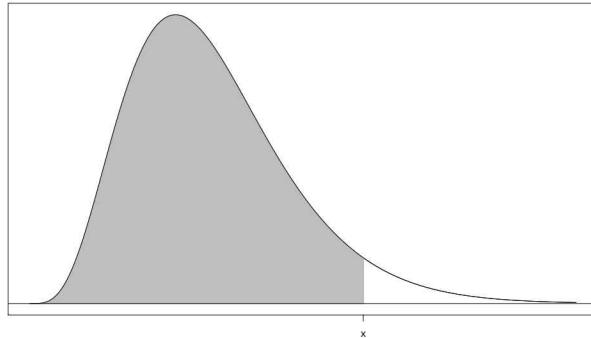
Lorsque la règle usuelle d'utilisation du test du khi-deux n'est pas satisfaite, on pourra faire des regroupements de modalités si cela a du sens ou utiliser le test exact de Fisher pour tester l'indépendance de deux variables notamment dans le cas où elles ont deux modalités chacunes.

## Loi du $\chi^2$

Soit une variable aléatoire  $KH$  distribuée suivant une loi du  $\chi_p^2$ . On considère la probabilité de trouver une valeur inférieure à  $x$  :  $F(x) = P(KH < x)$ .

Sur le graphique, cette probabilité correspond à l'aire en grisé.

Dans la table il y a en ligne les degrés de liberté (ddl), en colonne la probabilité  $F(x) = P(KH < x)$ , et à l'intersection la valeur de  $x$  correspondante.



	<b>0.001</b>	<b>0.005</b>	<b>0.010</b>	<b>0.025</b>	<b>0.05</b>	<b>0.1000</b>	<b>0.5000</b>	<b>0.9000</b>	<b>0.9500</b>	<b>0.9750</b>	<b>0.9900</b>	<b>0.9950</b>	<b>0.9990</b>
<b>1</b>	0.000	0.000	0.000	0.001	0.004	0.016	0.455	2.706	3.841	5.024	6.635	7.879	10.828
<b>2</b>	0.002	0.010	0.020	0.051	0.103	0.211	1.386	4.605	5.991	7.378	9.210	10.597	13.816
<b>3</b>	0.024	0.072	0.115	0.216	0.352	0.584	2.366	6.251	7.815	9.348	11.345	12.838	16.266
<b>4</b>	0.091	0.207	0.297	0.484	0.711	1.064	3.357	7.779	9.488	11.143	13.277	14.860	18.467
<b>5</b>	0.210	0.412	0.554	0.831	1.145	1.610	4.351	9.236	11.070	12.833	15.086	16.750	20.515
<b>6</b>	0.381	0.676	0.872	1.237	1.635	2.204	5.348	10.645	12.592	14.449	16.812	18.548	22.458
<b>7</b>	0.598	0.989	1.239	1.690	2.167	2.833	6.346	12.017	14.067	16.013	18.475	20.278	24.322
<b>8</b>	0.857	1.344	1.646	2.180	2.733	3.490	7.344	13.362	15.507	17.535	20.090	21.955	26.124
<b>9</b>	1.152	1.735	2.088	2.700	3.325	4.168	8.343	14.684	16.919	19.023	21.666	23.589	27.877
<b>10</b>	1.479	2.156	2.558	3.247	3.940	4.865	9.342	15.987	18.307	20.483	23.209	25.188	29.588
<b>11</b>	1.834	2.603	3.053	3.816	4.575	5.578	10.341	17.275	19.675	21.920	24.725	26.757	31.264
<b>12</b>	2.214	3.074	3.571	4.404	5.226	6.304	11.340	18.549	21.026	23.337	26.217	28.300	32.909
<b>13</b>	2.617	3.565	4.107	5.009	5.892	7.042	12.340	19.812	22.362	24.736	27.688	29.819	34.528
<b>14</b>	3.041	4.075	4.660	5.629	6.571	7.790	13.339	21.064	23.685	26.119	29.141	31.319	36.123
<b>15</b>	3.483	4.601	5.229	6.262	7.261	8.547	14.339	22.307	24.996	27.488	30.578	32.801	37.697
<b>16</b>	3.942	5.142	5.812	6.908	7.962	9.312	15.338	23.542	26.296	28.845	32.000	34.267	39.252
<b>17</b>	4.416	5.697	6.408	7.564	8.672	10.085	16.338	24.769	27.587	30.191	33.409	35.718	40.790
<b>18</b>	4.905	6.265	7.015	8.231	9.390	10.865	17.338	25.989	28.869	31.526	34.805	37.156	42.312
<b>19</b>	5.407	6.844	7.633	8.907	10.117	11.651	18.338	27.204	30.144	32.852	36.191	38.582	43.820
<b>20</b>	5.921	7.434	8.260	9.591	10.851	12.443	19.337	28.412	31.410	34.170	37.566	39.997	45.315
<b>21</b>	6.447	8.034	8.897	10.283	11.591	13.240	20.337	29.615	32.671	35.479	38.932	41.401	46.797
<b>22</b>	6.983	8.643	9.542	10.982	12.338	14.041	21.337	30.813	33.924	36.781	40.289	42.796	48.268
<b>23</b>	7.529	9.260	10.196	11.689	13.091	14.848	22.337	32.007	35.172	38.076	41.638	44.181	49.728
<b>24</b>	8.085	9.886	10.856	12.401	13.848	15.659	23.337	33.196	36.415	39.364	42.980	45.559	51.179
<b>25</b>	8.649	10.520	11.524	13.120	14.611	16.473	24.337	34.382	37.652	40.646	44.314	46.928	52.620
<b>26</b>	9.222	11.160	12.198	13.844	15.379	17.292	25.336	35.563	38.885	41.923	45.642	48.290	54.052
<b>27</b>	9.803	11.808	12.879	14.573	16.151	18.114	26.336	36.741	40.113	43.195	46.963	49.645	55.476
<b>28</b>	10.391	12.461	13.565	15.308	16.928	18.939	27.336	37.916	41.337	44.461	48.278	50.993	56.892
<b>29</b>	10.986	13.121	14.256	16.047	17.708	19.768	28.336	39.087	42.557	45.722	49.588	52.336	58.301
<b>30</b>	11.588	13.787	14.953	16.791	18.493	20.599	29.336	40.256	43.773	46.979	50.892	53.672	59.703
<b>31</b>	12.196	14.458	15.655	17.539	19.281	21.434	30.336	41.422	44.985	48.232	52.191	55.003	61.098
<b>32</b>	12.811	15.134	16.362	18.291	20.072	22.271	31.336	42.585	46.194	49.480	53.486	56.328	62.487
<b>33</b>	13.431	15.815	17.074	19.047	20.867	23.110	32.336	43.745	47.400	50.725	54.776	57.648	63.870
<b>34</b>	14.057	16.501	17.789	19.806	21.664	23.952	33.336	44.903	48.602	51.966	56.061	58.964	65.247
<b>35</b>	14.688	17.192	18.509	20.569	22.465	24.797	34.336	46.059	49.802	53.203	57.342	60.275	66.619
<b>36</b>	15.324	17.887	19.233	21.336	23.269	25.643	35.336	47.212	50.998	54.437	58.619	61.581	67.985
<b>37</b>	15.965	18.586	19.960	22.106	24.075	26.492	36.336	48.363	52.192	55.668	59.893	62.883	69.346
<b>38</b>	16.611	19.289	20.691	22.878	24.884	27.343	37.335	49.513	53.384	56.896	61.162	64.181	70.703
<b>39</b>	17.262	19.996	21.426	23.654	25.695	28.196	38.335	50.660	54.572	58.120	62.428	65.476	72.055
<b>40</b>	17.916	20.707	22.164	24.433	26.509	29.051	39.335	51.805	55.758	59.342	63.691	66.766	73.402
<b>41</b>	18.575	21.421	22.906	25.215	27.326	29.907	40.335	52.949	56.942	60.561	64.950	68.053	74.745
<b>42</b>	19.239	22.138	23.650	25.999	28.144	30.765	41.335	54.090	58.124	61.777	66.206	69.336	76.084
<b>43</b>	19.906	22.859	24.398	26.785	28.965	31.625	42.335	55.230	59.304	62.990	67.459	70.616	77.419
<b>44</b>	20.576	23.584	25.148	27.575	29.787	32.487	43.335	56.369	60.481	64.201	68.710	71.893	78.750
<b>45</b>	21.251	24.311	25.901	28.366	30.612	33.350	44.335	57.505	61.656	65.410	69.957	73.166	80.077
<b>46</b>	21.929	25.041	26.657	29.160	31.439	34.215	45.335	58.641	62.830	66.617	71.201	74.437	81.400
<b>47</b>	22.610	25.775	27.416	29.956	32.268	35.081	46.335	59.774	64.001	67.821	72.443	75.704	82.720
<b>48</b>	23.295	26.511	28.177	30.755	33.098	35.949	47.335	60.907	65.171	69.023	73.683	76.969	84.037
<b>49</b>	23.983	27.249	28.941	31.555	33.930	36.818	48.335	62.038	66.339	70.222	74.919	78.231	85.351
<b>50</b>	24.674	27.991	29.707	32.357	34.764	37.689	49.335	63.167	67.505	71.420	76.154	79.490	86.661

**Exemple d'utilisation :** pour  $n = 1$  et  $F(x) = 0.9500$  on trouve  $x = 3.84$ .

### 3 Principes de l'AFC

Rappelons que notre tableau de données est  $N$ , tableau de contingence à  $I$  lignes et  $J$  colonnes. Si on note  $D_I$  et  $D_J$  les matrices diagonales des effectifs marginaux des variables  $X$  et  $Y$  :

$$D_I = \text{diag}(n_{1.}, \dots, n_{I.}), \quad D_J = \text{diag}(n_{.1}, \dots, n_{.J}),$$

les tableaux des profils lignes et des profils-colonnes s'écrivent :

$$T_{pl} = D_I^{-1}N \quad \text{et} \quad T_{pc} = \boxed{\quad}.$$

Deux approches sont envisageables : l'analyse des profils lignes et l'analyse des profils colonnes.

#### 3.1 Analyse du nuage des profils-lignes

Si on s'intéresse aux lignes, on peut considérer le tableau  $T_{pl}$  des profils-lignes comme un tableau individus-caractères particulier et effectuer une ACP. Les "individus" de cette analyse sont les profils-lignes munis des poids

$$f_{1.} = \frac{n_{1.}}{n}, \dots, f_{I.} = \frac{n_{I.}}{n}.$$

Par analogie avec le chapitre sur l'ACP de l'enseignement *Statistique exploratoire multivariée* nous avons ici :

- $X_{(I,J)} = T_{pl} = D_I^{-1}N$
- $l_i = {}^t \left( \frac{n_{i1}}{n_{..}}, \dots, \frac{n_{iJ}}{n_{..}} \right) = {}^t (f_i^1, \dots, f_i^J) \in \mathbb{R}^J$ .
- $\mathbb{R}^J$  = espace des profils-lignes ("individus").

Chaque profil-ligne  $l_i$  est muni du poids  $f_{i.} = \frac{n_{i.}}{n}$ , correspondant à sa fréquence, et reflétant l'importance de la modalité  $y_i$  par rapport aux autres modalités de  $Y$ .

**Définition 3.1** On appelle **nuage des profils-lignes**  $\mathcal{M}_l$ , l'ensemble des  $I$  points  $l_i$  de  $\mathbb{R}^J$  munis de leurs poids  $f_{i.}$  :

$$\mathcal{M}_l = \{(l_i, f_{i.}) ; i = 1, \dots, I\}.$$

**Propriétés :**

1. Le centre de gravité  $G_X$  du nuage  $\mathcal{M}_l$  (profil-moyen des lignes) a pour coordonnées :

$$G_X = \boxed{\quad}$$

$$= \boxed{\quad}$$

2. Les profils ayant pour somme 1, les points  $l_i$  de  $\mathcal{M}_l$  ainsi que leur centre de gravité appartiennent à un sous-espace affine de  $\mathbb{R}^J$  l'hyperplan  $\mathcal{H}_{J-1}$  de dimension  $J - 1$  défini par :

$$\mathcal{H}_{J-1} = \left\{ (x_1, \dots, x_J) \in \mathbb{R}^J ; \sum_{i=1}^J x_i = 1 \right\}.$$

### 3.2 Analyse du nuage des profils-colonnes

Ici les "individus" sont les profils-colonnes munis des poids

$$f_{\cdot 1} = \frac{n_{\cdot 1}}{n}, \dots, f_{\cdot J} = \frac{n_{\cdot J}}{n}.$$

- $X_{(I,J)} = T_{pc} = ND_J^{-1}$
  - profil-colonne :  $c_j = {}^t \left( \frac{n_{1j}}{n_{..j}}, \dots, \frac{n_{Ij}}{n_{..j}} \right) \in \mathbb{R}^I$ .
  - $\mathbb{R}^I$  = espace des profils-colonnes ("individus")

**Définition 3.2** On appelle nuage des profils-colonnes  $\mathcal{M}_c$ , l'ensemble des  $J$  points  $c_j$  de  $\mathbb{R}^I$  munis de leurs poids  $f_{.j}$  :

$$\mathcal{M}_c = \{(c_j, f_{\cdot j}) ; j = 1, \dots, J\}.$$

## Propriétés :

- Le centre de gravité  $G_Y$  du nuage  $\mathcal{M}_c$  (profil-moyen des colonnes) a pour coordonnées :  

$$G_Y = {}^t\left(\frac{n_1}{n}, \dots, \frac{n_I}{n}\right) \in \mathbb{R}^I$$

$$= {}^t(f_1, f_2, \dots, f_I)$$
  - Les profils ayant pour somme 1, les points  $c_j$  de  $\mathcal{M}_c$  ainsi que leur centre de gravité appartiennent à un sous-espace affine de  $\mathbb{R}^I$  l'hyperplan  $\mathcal{H}_{I-1}$  de dimension  $I - 1$  défini par :

$$\mathcal{H}_{I-1} = \left\{ (x_1, \dots, x_I) \in \mathbb{R}^I ; \sum_{i=1}^I x_i = 1 \right\}.$$

Dans le cas de l'indépendance statistique

$$n_{ij} = \boxed{\phantom{000}},$$

c'est à dire

$$(pl)_j = \frac{n_{ij}}{n_i} = \boxed{\phantom{000}} \quad \text{et} \quad \frac{n_{ij}}{n_{\cdot j}} = \boxed{\phantom{000}} = (pc)_i.$$

Les nuages  $\mathcal{M}_l$  et  $\mathcal{M}_c$  sont donc réduits à l'ensemble des points  $x$  tels que

Ainsi dans le cas général, l'étude de la forme de ces nuages au moyen d'ACP permet de rendre compte

Maintenant se pose la question suivante : quelle métrique choisir pour chacun de ces espaces ?

### 3.3 Métrique du khi2

Distance entre deux profils-lignes  $l_i$  et  $l_{i'}$  :

$$\begin{aligned}
d_{\chi^2}^2(l_i, l_{i'}) &= \sum_{j=1}^J \frac{n}{n.j} \left( \frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2 \\
&= \sum_{j=1}^J \frac{\left( \frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2}{\frac{n.j}{n}} \\
&\equiv t(l_i - l_{i'}) M(l_i - l_{i'}) \equiv l_i - l_{i'}, l_i - l_{i'} \geq M
\end{aligned}$$

avec  $M = nD_L^{-1}$ , métrique diagonale.

### Justification de l'utilisation de cette métrique :

1. La pondération par  $\frac{n}{n_{.j}}$  de chaque carré de différence revient à donner des importances comparables aux diverses modalités  $y_j$ . Sinon sans pondération, la distance reflète surtout la différence entre les modalités de plus grands effectifs.  
 → Considérant l'exemple des couleurs, calculer la distance entre les deux premières lignes en considérant la distance euclidienne (somme des carrés des différences de pourcentages des divers diplômes), puis la distance du  $\chi^2$ . Interpréter les différences obtenues.

#### 2. Un autre argument très important :

Supposons que deux colonnes de  $N$ ,  $c_j$  et  $c_{j'}$  ont même profil c'est à dire

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{ij'}}{n_{.j'}} \text{ pour tout } i = 1, \dots, I$$

il est logique de les regrouper en une seule colonne d'effectif  $n_{ij} + n_{ij'}$ . Cependant cette opération ne doit pas modifier les distances entre profils-lignes. Il se trouve que la distance du khi-deux vérifie cette propriété. En effet on montre que

$$\frac{n}{n_{.j}} \left( \frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2 + \frac{n}{n_{.j'}} \left( \frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}} \right)^2 = \frac{n}{n_{.j} + n_{.j'}} \left( \frac{n_{ij} + n_{ij'}}{n_{i.}} - \frac{n_{i'j} + n_{i'j'}}{n_{i'.}} \right)^2.$$

Ceci n'est pas vérifié par la métrique euclidienne. On a un résultat analogue pour l'agrégation de deux profils lignes identiques.

**En résumé :** L'agrégation de deux modalités d'une même variable ayant des profils identiques en une nouvelle modalité affecté de la somme de leurs effectifs ne change rien

- ni aux distances entre les modalités de cette variable
- ni aux distances entre les modalités de l'autre variable

Cette propriété fondamentale est appelée **la propriété d'équivalence distributionnelle**. Elle garantit une certaine invariance par rapport à la nomenclature choisie pour la construction des modalités d'une variable. On ne perd pas d'information en agrégeant les modalités de mêmes profils et l'on n'en gagne pas en subdivisant.

### Distance entre profils-colonnes :

$$\begin{aligned} d_{\chi^2}^2(c_j, c_{j'}) &= \boxed{\quad} \\ &= \boxed{\quad}, \\ \text{avec } M &= nD_I^{-1}, \text{ métrique diagonale.} \end{aligned}$$

La distance du khi et le rôle symétrique joué par les lignes et les colonnes du tableau caractérisent l'AFC et lui assurent des propriétés remarquables que ne possède pas l'ACP : équivalence distributionnelle et les relations quasi-barycentriques (voir plus loin).

### 3.4 Inertie des nuages

**Propriété 3.1**  $I(\mathcal{M}_l) = I(\mathcal{M}_c) = \frac{1}{n} \chi^2$

$$\text{Dém : } I(\mathcal{M}_l) = \sum_{i=1}^I f_i d_{\chi^2}^2(l_i, G_X)$$

$$\begin{aligned}
 &= \boxed{\quad} \\
 &= \boxed{\quad} \\
 &= \frac{1}{n} \chi^2.
 \end{aligned}$$

$$\rightarrow \text{Montrer de même que } I(\mathcal{M}_c) = \frac{1}{n} \chi^2$$

$\rightarrow$  Comment s'interprète l'inertie de chaque nuage ?

### 3.5 ACP des deux nuages profils

Nous avons défini deux nuages de points : les profils lignes et les profils colonnes munis de leurs poids et de la métrique du  $\chi^2$ . Deux ACP, chacune associée à un nuage, sont alors possibles : elles définissent l'analyse factorielle des correspondances associée au tableau de contingence  $N$  :

1) ACP du nuage des profils-lignes avec

- Tableau de données  $X(I, J) : \mathcal{X} = D_I^{-1} N$
- Espace des individus :  $\mathbb{R}^J$
- Métrique :  $M = n D_J^{-1}$
- Poids :  $P = \frac{1}{n} D_I$

2) ACP du nuage des profils-colonnes avec

- Tableau de données  $X(J, I) : \mathcal{Y} = D_J^{-1} {}^t N$
- Espace des individus :  $\mathbb{R}^I$
- Métrique :  $M = n D_I^{-1}$
- Poids :  $P = \frac{1}{n} D_J$

Les résultats sont en dualité exacte. En AFC les lignes et les colonnes ont un rôle symétrique.

#### Matrice de variance-covariance

$$V = {}^t \mathcal{X} P \mathcal{X} - G_X^t G_X = {}^t \tilde{\mathcal{X}} P \tilde{\mathcal{X}}$$

Or on montre que la droite  $(OG_X)$  est orthogonale au sens de la métrique du khi2 au sous-espace contenant le nuage. Par exemple pour le nuage ligne : soit  $x \in \mathcal{H}_{J-1}$ ,

$$\langle OG_X, G_X x \rangle_{\chi^2} = {}^t(x - G_X) M G_X,$$

$$\text{avec } MG_X = n D_J^{-1} G_X = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \text{ D'où } {}^t(x_l - G_X) M G_X = \sum_{j=1}^s (x_l(j) - G_X(j)) = 1 - 1 = 0.$$

En conclusion le vecteur  $OG_X$  est orthogonal au support du nuage, il est donc vecteur propre de  $VM$  associé à la valeur propre  $\lambda = 0$ .

**Remarque 3.1**    1. Il y a au plus  $\min(I-1, J-1)$  valeurs propres non nulles, donc au plus  $\min(I-1, J-1)$  axes principaux dans une AFC.

2. On montre que les vecteurs propres de  $VM$  associés aux valeurs propres non nulles sont les mêmes que ceux de la matrice  ${}^t\mathcal{X}P\mathcal{X}M$  associés aux mêmes valeurs propres. Par contre  $G_X$  est vecteur propre associé à la vp 0 pour  $VM$  et associé à la vp 1 pour la matrice  ${}^t\mathcal{X}D\mathcal{X}M$ .

D'après le point 2 de la remarque précédente, il est donc inutile de centrer les tableaux de profils. De manière pratique, on effectuera donc des ACP non centrées. Puis à la vp 1 associée à l'axe ( $OG_X$ ), on fera correspondre la vp 0 pour se ramener à une étude centrée en l'origine.

**Tableau récapitulatif :**

Nuage $\mathcal{M}_l$ des $I$ profils-lignes dans $\mathbb{R}^J$	Eléments de base	Nuage $\mathcal{M}_c$ des $J$ profils-colonnes dans $\mathbb{R}^I$
$T_{pl} = D_I^{-1}N$ ( $I, J$ )	Tableau de données	${}^tT_{pc} = D_J^{-1}{}^tN$ ( $J, I$ )
$L = {}^tT_{pl}D_IT_{pl}D_J^{-1}$ ( $J, J$ )	Matrice à diagonaliser	$C = T_{pc}D_J^tT_{pl}D_I^{-1}$ ( $I, I$ )
$u_k = {}^t(u_{1k}, \dots, u_{Jk})$ $Lu_k = \lambda_k u_k$ avec $\lambda_k \neq 1$	Vecteurs (axes principaux)	$v_k = {}^t(v_{1k}, \dots, v_{Ik})$ $Cv_k = \lambda_k v_k$ avec $\lambda_k \neq 1$
Coordonnées des $I$ profils-lignes $l_i$ sur $\Delta_{u_k}$ (projections sur $\Delta_{u_k}$ )  $c_k(i) = n \times \sum_{l=1}^J \frac{n_{il}}{n_i \cdot n_l} u_{lk}$ $c_k = n \times D_I^{-1}ND_J^{-1}u_k = n \sqrt{\lambda_k}D_I^{-1}v_k$	Coordonnées factorielles (facteurs principaux)	Vecteurs propres de $L$ normalisés par $\lambda$ Coordonnées de la projection de $c_j$ sur $\Delta_{v_k}$ $d_k(j) = n \times \sum_{l=1}^I \frac{n_{lj}}{n_l \cdot n_j} v_{lk}$ $d_k = n \times D_J^{-1}{}^tND_I^{-1}v_k = n \sqrt{\lambda_k}D_J^{-1}u_k$

### Remarque 3.2

1. Les facteurs principaux  $c_k$  (resp.  $d_k$ ) sont orthogonaux deux à deux, centrés et de variance égale à  $\lambda_k$ .
2. Pour tout axe  $\Delta_k$  le barycentre  $G_X$  se projette en  $O$ .
3. L'inertie du nuage projeté sur le  $k$ ème axe est  $\lambda_k$ .

### 3.6 Relations de transition

Les coordonnées des profils-lignes et celles des profils-colonnes sont reliées par des formules simples dont un des intérêts principaux est d'éviter deux diagonalisations. On diagonalisera la matrice de plus petite dimension.

On peut montrer que les matrices  $L$  et  $C$  ont mêmes valeurs propres non nulles  $\lambda_k$ . Leurs vecteurs propres sont reliés par les relations suivantes :

**Théorème 3.1** Soit  $p = \text{rang}(L) = \text{rang}(C)$ . Pour tout  $k \leq p$ , il existe une relation dite de transition, entre les vecteurs propres  $u_k$  et  $v_k$  :

$$\begin{aligned} v_k &= \frac{1}{\sqrt{\lambda_k}} T_{pc} u_k = \frac{1}{\sqrt{\lambda_k}} N D_J^{-1} u_k, \\ u_k &= \frac{1}{\sqrt{\lambda_k}} {}^t T_{pl} v_k = \frac{1}{\sqrt{\lambda_k}} {}^t N D_I^{-1} v_k \end{aligned}$$

Le résultat précédent conduit aux relations fondamentales de l'AFC reliant les coordonnées factorielles entre elles, dites **les relations quasi-barycentriques** :

**Théorème 3.2** Soit  $p = \text{rang}(L) = \text{rang}(C)$ . Pour tout  $k \leq p$ ,

$$\begin{aligned} c_k(i) &= \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^J \frac{n_{ij}}{n_{i \cdot}} d_k(j), \\ d_k(j) &= \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^I \frac{n_{ij}}{n_{ \cdot j}} c_k(i). \end{aligned}$$

**Interprétation :**

A un coefficient de dilatation près de  $\frac{1}{\sqrt{\lambda_k}}$ , la projection du profil-ligne  $i$  sur l'axe  $k$  est le barycentre des projections des profils-colonnes pondérés par les fréquences conditionnelles du profil  $i$  (part de la modalité  $j$  sachant  $i$ ). Et on a une interprétation analogue pour la projection du profil colonne  $j$ .

La parfaite symétrie entre ACP profil-ligne et ACP profil-colonne (relations quasi-barycentriques) conduisent à superposer les plans principaux des deux ACP afin d'obtenir une représentation simultanée des deux nuages. Mais attention les distances entre les projections des profils-lignes et des profil-colonnes n'ont pas de sens puisque ces points sont dans des espaces différents. On ne peut les interpréter qu'en termes de quasi-barycentres.

→ En utilisant les relations quasi-barycentriques, montrer que les valeurs propres  $\lambda_k$  de l'AFC sont inférieures ou égales à 1.

### 3.7 Formules de reconstitution

Comme en ACP on dispose de formules dites de reconstitution permettant de reconstituer le tableau  $N$  à partir des coordonnées factorielles  $c_k$  et  $d_k$  : Pour tout  $i \leq I$  et tout  $j \leq J$ , on a :

$$n_{ij} = \frac{n_{i \cdot} n_{ \cdot j}}{n} \left( 1 + \sum_{k=1}^p \frac{1}{\sqrt{\lambda_k}} c_k(i) d_k(j) \right).$$

Les facteurs et vp expliquent donc en quoi les  $n_{ij}$  s'écartent des  $\frac{n_{i \cdot} n_{ \cdot j}}{n}$  (effectif attendu sous l'hypothèse d'indépendance).

## 4 Aides à l'interprétation

Rappelons que deux types de coefficients apportent de l'information intéressante pour l'interprétation des plans factoriels.

- **Contribution relative** : exprime la part prise par une modalité de la variable dans l'inertie "expliquée" par un facteur.
- **Cosinus carré** : mesure la qualité de représentation de la modalité sur le facteur. Il exprime en fait la part prise par un facteur dans la dispersion de la modalité d'une variable.

**Contribution relative des profils à l'inertie de l'axe  $k$  :**

$$\begin{aligned} CTR_k(i) &= \frac{f_{i\cdot}c_k^2(i)}{\lambda_k} \quad \text{avec } f_{i\cdot} = n_{i\cdot}/n \text{ le poids de } l_i. \\ CTR_k(j) &= \frac{f_{\cdot j}d_k^2(j)}{\lambda_k} \quad \text{avec } f_{\cdot j} = n_{\cdot j}/n \text{ le poids de } c_j. \end{aligned}$$

**Qualité de représentation sur l'axe  $k$  :**

$$\begin{aligned} QLT_k(i) &= \frac{c_k^2(i)}{\sum_{k=1}^p c_k^2(i)} \\ QLT_k(j) &= \frac{d_k^2(j)}{\sum_{k=1}^p d_k^2(j)} \end{aligned}$$

**Eléments supplémentaires :**

Les points supplémentaires sont des profils qui n'entrent pas dans la construction des axes mais qui sont représentés dans les plans factoriels. Leur position est interprétée en terme de quasi-barycentre.

Si on dispose par exemple d'une modalité supplémentaire  $x_0$  de la variable  $X$ , on peut calculer la coordonnée du profil-ligne correspondant  $l_0$  sur l'axe  $k$  en utilisant les formules quasi-barycentriques :

$$c_k(0) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^J \frac{n_{0j}}{n_0} d_k(j),$$

où  $n_0$  est l'effectif total de la modalité supplémentaire et  $n_{0j}$  l'effectif du croisement  $(x_0, y_j)$ . D'autre part on peut également évaluer la qualité de représentation des points supplémentaires sur l'axe  $k$  par le calcul du  $\cos^2$ .

## 5 Interpréter les résultats d'une AFC

L'interprétation des résultats d'une AFC suit généralement les étapes suivantes.

1. Vérifier l'existence d'une dépendance entre les deux variables à l'aide du test du chi-deux.
2. Sélectionner les dimensions des sous espaces considérés expliquant les nuages.
3. Énoncer le taux d'inertie expliqué par le sous-espace et indiquer le taux expliqué par chacun des axes
4. Étudier la qualité de représentation des modalités des profils-lignes et colonnes et exclure de l'analyse les modalités mal représentées
5. Interprétation des axes

6. Étude du nuage des profils-lignes et du nuage des profils-colonnes :

- Les proximités des modalités d'une même variable par exemple les profils-lignes (resp. les profils-colonnes) bien représentées dans un plan, peuvent être interprétées en termes de similitude de profils.
- Les proximités des modalités de variables différentes (proximité d'un profil-ligne et d'un profil-colonne) est plus « périlleuse » : les profils-lignes et les profils-colonnes n'appartiennent pas au même espace. Même si ces modalités sont bien représentées, on ne peut seulement les interpréter qu'en termes de barycentre : les individus possédant la modalité  $i$  de la première variable ont un barycentre proche de celui des individus possédant la modalité  $j$ . Cette proximité peut révéler un trait caractéristique mais attention ceci n'est pas toujours le cas. Pour s'en assurer il suffit d'examiner les tableaux des profils-lignes et des profils-colonnes de départ. Une proximité de modalités de variables différentes doit être considérée comme une conjecture de trait caractéristique qui doit être validée ou infirmée en examinant les tableaux de départ.

7. Faire un petit résumé en dégageant si possible une ou des hypothèses de travail.

→ A partir du listing présentant ci-dessous les sorties SAS de l'AFC sur le jeu de données couleurs, proposer une interprétation des résultats obtenus pour l'exemple des couleurs des yeux et cheveux.

## La procédure CORRESP

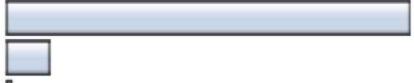
Table de contingence					
	brun	chatain	roux	blond	Somme
<b>marron</b>	68	119	26	7	220
<b>noisette</b>	15	54	14	10	93
<b>vert</b>	5	29	14	16	64
<b>bleu</b>	20	84	17	94	215
<b>Somme</b>	108	286	71	127	592

Profils de lignes				
	brun	chatain	roux	blond
<b>marron</b>	0.309091	0.540909	0.118182	0.031818
<b>noisette</b>	0.161290	0.580645	0.150538	0.107527
<b>vert</b>	0.078125	0.453125	0.218750	0.250000
<b>bleu</b>	0.093023	0.390698	0.079070	0.437209

Profils de colonnes				
	brun	chatain	roux	blond
<b>marron</b>	0.629630	0.416084	0.366197	0.055118
<b>noisette</b>	0.138889	0.188811	0.197183	0.078740
<b>vert</b>	0.046296	0.101399	0.197183	0.125984
<b>bleu</b>	0.185185	0.293706	0.239437	0.740157

# Le Système SAS

## La procédure CORRESP

<b>Décomposition de l'inertie et du Khi-2</b>							
<b>Valeur singulière</b>	<b>Inertie principale</b>	<b>Khi-2</b>	<b>Pourcentage</b>	<b>Pourcentage cumulé</b>			
					0	20	40
0.45692	0.20877	123.593	89.37	89.37			
0.14909	0.02223	13.158	9.51	98.89		20	
0.05097	0.00260	1.538	1.11	100.00		40	60
	0.23360	138.290	100.00				80
							Degrés de liberté = 9

<b>Coordonnées des lignes</b>		
	<b>Dim1</b>	<b>Dim2</b>
<b>marron</b>	-0.4922	-0.0883
<b>noisette</b>	-0.2126	0.1674
<b>vert</b>	0.1618	0.3390
<b>bleu</b>	0.5474	-0.0830

<b>Statistiques descriptives pour les points des lignes</b>			
	<b>Qualité</b>	<b>Masse</b>	<b>Inertie</b>
<b>marron</b>	0.9981	0.3716	0.3985
<b>noisette</b>	0.8787	0.1571	0.0560
<b>vert</b>	0.9484	0.1081	0.0689
<b>bleu</b>	0.9999	0.3632	0.4766

<b>Contributions partielles à l'inertie des points des lignes</b>		
	<b>Dim1</b>	<b>Dim2</b>
<b>marron</b>	0.4312	0.1304
<b>noisette</b>	0.0340	0.1980
<b>vert</b>	0.0135	0.5591
<b>bleu</b>	0.5213	0.1124

<b>Indices des coordonnées qui contribuent le plus à l'inertie des points de ligne</b>			
	<b>Dim1</b>	<b>Dim2</b>	<b>Meilleur</b>
<b>marron</b>	1	1	1
<b>noisette</b>	0	2	2
<b>vert</b>	0	2	2
<b>bleu</b>	1	0	1

## Le Système SAS

### La procédure CORRESP

Carrés du cosinus pour les points des lignes		
	Dim1	Dim2
<b>marron</b>	0.9670	0.0311
<b>noisette</b>	0.5424	0.3363
<b>vert</b>	0.1759	0.7726
<b>bleu</b>	0.9775	0.0224

Coordonnées des colonnes		
	Dim1	Dim2
<b>brun</b>	-0.5046	-0.2148
<b>chatain</b>	-0.1483	0.0327
<b>roux</b>	-0.1295	0.3196
<b>blond</b>	0.8353	-0.0696

Statistiques descriptives pour les points des colonnes			
	Qualité	Masse	Inertie
<b>brun</b>	0.9899	0.1824	0.2373
<b>chatain</b>	0.9063	0.4831	0.0526
<b>roux</b>	0.9451	0.1199	0.0646
<b>blond</b>	0.9996	0.2145	0.6455

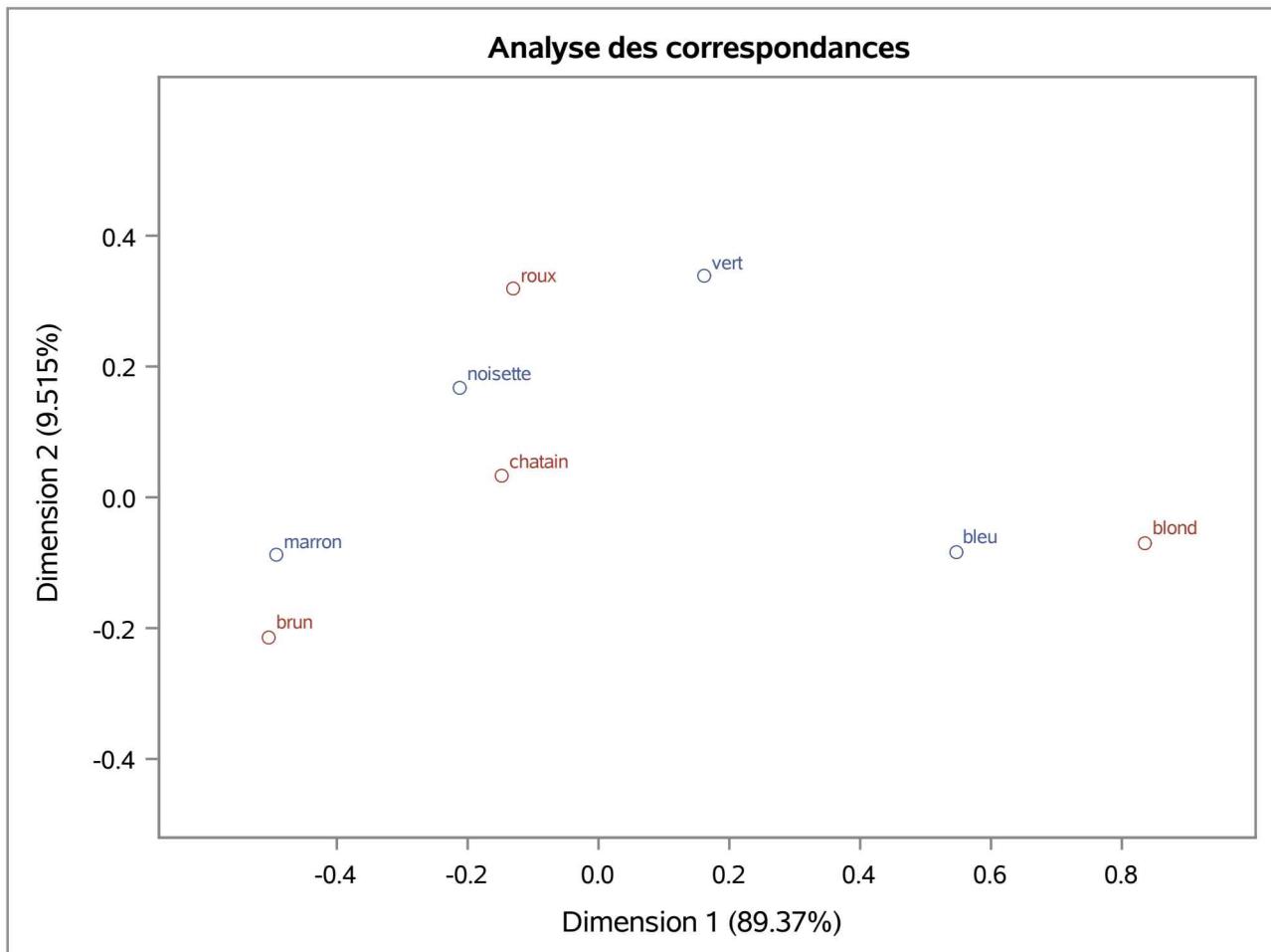
Contributions partielles à l'inertie des points des colonnes		
	Dim1	Dim2
<b>brun</b>	0.2225	0.3788
<b>chatain</b>	0.0509	0.0232
<b>roux</b>	0.0096	0.5513
<b>blond</b>	0.7170	0.0467

Indices des coordonnées qui contribuent le plus à l'inertie des points de colonne			
	Dim1	Dim2	Meilleur
<b>brun</b>	2	2	2
<b>chatain</b>	0	0	1
<b>roux</b>	0	2	2
<b>blond</b>	1	0	1

## La procédure CORRESP

Carrés du cosinus pour les points des colonnes		
	Dim1	Dim2
<b>brun</b>	0.8380	0.1519
<b>chatain</b>	0.8644	0.0420
<b>roux</b>	0.1333	0.8118
<b>blond</b>	0.9927	0.0069

## La procédure CORRESP



## 6 Références

L. Lebart, M. Piron et A. Morineau (2006). *Statistique exploratoire multidimensionnelle*, Editions Dunod, Paris.

G. Saporta (2006). *Probabilités, Analyse des données et statistique*. Editions Technip, Paris.